ISS-SP-070

ＲＮＮを用いたＦ０操作による歌唱音声の熱唱化の検討

早坂　琢真†　　伊藤　彰則††
† 東北大学工学部　　†† 東北大学大学院工学研究科

1. はじめに

歌唱音声の変換技術として，地声-裏声の変換[1]や，奄美大島の民謡で多用される「グイン」と呼ばれる表現の歌声への付加[2]などの研究が既になされてきたが，本稿では熱唱度という概念に着目し，歌唱音声を熱唱化するということを，RNN を用いて検討する．

2. 熱唱度とは

本稿における熱唱度とは，人がいかに一生懸命に歌唱しているかを表す度合いである．大道らの研究[3],[4]において，「歌唱者本人自身の評価による熱唱度」と「歌唱音声を聞いた第 3 者が知覚する熱唱度」とは区別してあるが，本稿でもそれらを区別して扱う．

3. 熱唱化手法

3.1. 学習モデル

F0 の学習に用いるエルマンネット（Elman Network）は再帰型ニューラルネット（RNN，Recurrent Neural Network）の一種である．エルマンネットの入力層は現在の時刻の入力ユニットと直前の時刻の中間層のコピーから成る文脈ユニットとで構成される．すなわちエルマンネットのネットワークの状態は現在の入力と過去の履歴とで定まる．

3.2. F0 の変換

非熱唱時の F0 系列を$F0_n$，熱唱時の F0 系列を$F0_e$と表した時に，

$$\phi(F0_n) \cong F0_e \qquad (1)$$

となるような関数ϕをエルマンネットで学習させる．ここで，$F0_n$と$F0_e$は動的計画法（DP）によるマッチングを用いて，各フレームの対応をとってある．教師データには，男性 11 人にそれぞれ好みのポピュラーソングを熱唱，非熱唱と 2 通りの歌唱をそれぞれ 2 度ずつ行ってもらい，サビ周辺から 4 か所（1 つ当たり 3~6 秒）を取り出し，各人熱唱・非熱唱の組が 8 組ずつとなるよう用意したものを用いる．なお，サンプリング周波数は 44.1kHz，量子化ビット数は 16bit である．

3.3. 学習設定

隠れ層のユニット数は 20, 最適化アルゴリズムには Rprop を選んだ. なお, 実装などには R を用いた.

4. 熱唱化結果

図 1 に非熱唱時の F0 系列と，構築したエルマンネットワークによりそれを熱唱化変換させた F0 系列を示す. F0 の変換後は非熱唱なものと比べ短周期の変動が目立つ．これは，ビブラート部分を学習した結果と思われるが，実際のビブラートより変動が小さく，周期が短い．また，変換後得られた歌唱音声を実際に聴いたが，もととの差をあまり感じられなかった．

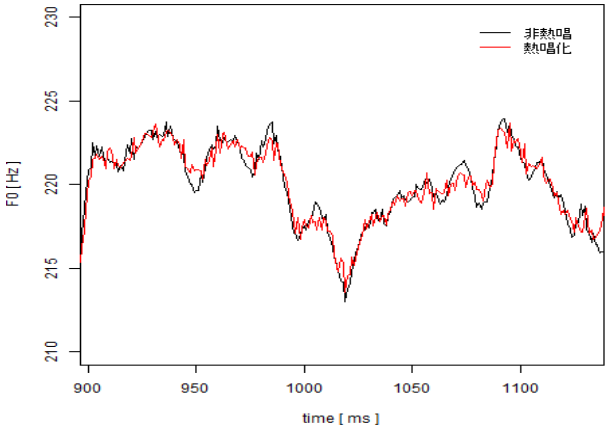

図 1：熱唱化結果

5. 今後の課題

ネットワークの内容（ユニット数など），LSTM を考慮するなどの見直しが必要である．

参考文献

[1] 森下ほか，信学技報，vol. 114，no. 358，pp. 83-87，2014
[2] 村主ほか，情処研報（MUS）, vol. 2010, no. 8, pp. 1-6, 2010.
[3] 大道ほか，情処研報（MUS），vol. 2010, no. 10, pp. 1-6, 2010
[4] 大道ほか，情処研報（MUS）vol. 2012, no. 2, pp. 1-6, 2012.